

# Learning to Recognize Affective Polarity in Similes

Ashequl Qadir and Ellen Riloff

School of Computing  
University of Utah  
Salt Lake City, UT 84112, USA  
{asheq, riloff}@cs.utah.edu

Marilyn A. Walker

Natural Language & Dialogue Systems Lab  
University of California Santa Cruz  
Santa Cruz, CA 95064, USA  
mawalker@ucsc.edu

## Abstract

A *simile* is a comparison between two essentially unlike things, such as “*Jane swims like a dolphin*”. Similes often express a positive or negative sentiment toward something, but recognizing the polarity of a simile can depend heavily on world knowledge. For example, “*memory like an elephant*” is positive, but “*memory like a sieve*” is negative. Our research explores methods to recognize the polarity of similes on Twitter. We train classifiers using lexical, semantic, and sentiment features, and experiment with both manually and automatically generated training data. Our approach yields good performance at identifying positive and negative similes, and substantially outperforms existing sentiment resources.

## 1 Introduction

A *simile* is a form of figurative language that compares two essentially unlike things (Paul, 1970), such as “*Jane swims like a dolphin*”. Similes often express a positive or negative view toward an entity, object, or experience (Li et al., 2012; Fishelov, 2007). Sometimes, the sentiment of a simile is expressed explicitly, such as “*Jane swims beautifully like a dolphin!*”. But in many cases the sentiment is implicit, evoked entirely from the comparison itself. “*Jane swims like a dolphin*” is easily understood to be a compliment toward Jane’s swimming ability because dolphins are known to be excellent swimmers.

A simile consists of four key components: the **topic** or **tenor** (subject of the comparison), the **vehicle** (object of the comparison), the **event** (act or state), and a **comparator** (usually “as”, “like”, or “than”) (Niculae and Danescu-Niculescu-Mizil, 2014). A **property** (shared attribute) can be optionally included as well (e.g., “*He is as red as*

*a tomato*”). Our research aims to identify the affective polarity of a simile as *positive*, *negative*, or *neutral*, based on its component phrases.

Table 1 shows examples of similes and their polarity. A simile can have *neutral* polarity if it offers an objective observation. Example (a) is a neutral simile because, although bananas have a distinctive smell, it is not generally considered to be a particularly good or bad scent. Example (b) illustrates that using the subjective verb “stink” instead of “smell” indicates a negative polarity toward the scent of bananas. Example (c) shows that including a subjective adjective such as “rotten” suggests a negative sentiment. Example (d) has negative polarity because the vehicle term, “garbage”, carries a strong negative connotation.

	Simile	Polarity
(a)	smells like bananas	<i>neutral</i>
(b)	stinks like bananas	<i>negative</i>
(c)	smells like rotten bananas	<i>negative</i>
(d)	smells like garbage	<i>negative</i>
(e)	memory like an elephant	<i>positive</i>
(f)	memory like a sieve	<i>negative</i>
(g)	looks like a celebrity	<i>positive</i>
(h)	acts like a celebrity	<i>negative</i>

Table 1: Simile Examples with Affective Polarity.

However, the affective polarity of a simile often emerges from multiple component terms. For instance, all of the words in Examples (e) and (f) have neutral polarity. But Example (e) is positive because elephants are widely known to have excellent memories, while Example (f) is negative because a sieve has holes, which is metaphorical with memory lapses. Examples (g) and (h) illustrate that a prior connotation can even be overridden depending upon the property being compared. In general, the word “celebrity” tends to have a positive connotation and looking like a celebrity is generally a compliment. But acting like a celebrity is a negative simile because it alludes to negative attributes such as narcissism or entitlement.

Our research explores the challenge of identifying the affective polarity of similes. First, we introduce a new data set of similes extracted from Twitter. We describe a manual annotation study to label them with affective polarity. We also present several approaches for identifying some instances of positive and negative similes using existing sentiment resources, to automatically create labeled data to train a classifier. Second, we describe a machine learning classifier to recognize the affective polarity of similes by considering lexical, semantic, and sentiment properties of their components. Third, we present experimental results for the simile polarity classifier, using both manually annotated training data and automatically labeled training data. Our evaluation shows that the classifier trained with manually labeled data achieves good performance at identifying positive and negative similes. Training with automatically labeled data produces classifiers that are not quite as good as those trained with manually labeled data, but they still substantially outperform existing sentiment resources and offer a way to easily train simile classifiers for different domains.

## 2 Related Work

Although similes are a popular form of comparison, there has been relatively little prior research on understanding affective polarity in similes. Veale and Hao (2007) created a large simile case-base using the pattern “as ADJ as a/an NOUN”. They collected similes by querying the web after instantiating part of the pattern with adjectives, and then had a human annotate 30,991 of the extracted similes for validity. Their focus was on extracting salient properties associated with simile **vehicles**, and the affective perception on vehicles that the salient properties bring about.

Veale (2012) took a step further and automatically recognized the affect toward vehicles when properties reinforce each other (e.g., hot and humid). They built a support graph of properties and determined how they connect to unambiguous positive and negative words. Li et al. (2012) used similar patterns to retrieve similes and determine basic sentiment toward simile vehicles across different languages using the compared properties. One major difference with their work and ours is that they determine sentiment or affective perception toward entities or concepts extracted from simile **vehicles**. In contrast, our work is focused

on determining affective polarity of a simile as a whole, where the affective polarity typically relates to an act or state of the **tenor**. In many cases, a simile vehicle does not have positive or negative polarity by itself. For example, “sauna” is not a positive or negative concept, but “room feels like a sauna” is a negative simile because it suggests that the room is humid and unpleasant.

Niculesae and Danescu-Niculescu-Mizil (2014) created a simile data set from Amazon product reviews, and determined when comparisons are figurative. They did not identify affective polarity, but showed that sentiment and figurative comparisons are correlated. Fishelov (2007) conducted a study of 16 similes where the connection between tenor and vehicle is obvious or not obvious, and when a conventional or unconventional explicit property is present or absent. Fishelov analyzed responses from participants to understand the positive and negative impression a simile conveys toward its tenor. Hanks (2005) presented an analysis of semantic categories of simile vehicles (animal, roles in society, artifact, etc.) that people most commonly use in similes.

Previous research has also explored sentiment expressed through metaphor. Rumbell et al. (2008) presented an analysis of animals that are metaphorically used to describe a person. Rentoumi et al. (2009) determined use of figurative language by disambiguating word senses, and then determined sentiment polarity at the sense level using ngram graph similarity. Wallington et al. (2011) identified affect in metaphor and similes when a comparison is made with an animal (e.g., dog, fox) or mythical creature (e.g., dragon, angel) by analyzing WordNet sense glosses of the compared terms. More recently, the SemEval-2015 Shared Task 11 (Ghosh et al., 2015) has addressed the sentiment analysis of figurative language such as irony, metaphor and sarcasm in Twitter.

Our work is also related to sentiment analysis in general. The most common approach applies supervised classification with features such as ngrams, parts-of-speech, punctuation, lexicon features, etc. (e.g., (Kouloumpis et al., 2011; Davidov et al., 2010; Mohammad et al., 2013)). To overcome the challenge of acquiring manually labeled data, some work automatically collects noisy training data using emoticons and hashtags (e.g., (Go et al., 2009; Purver and Battersby, 2012)). In addition to determining overall sen-

timent, research has also focused on understanding people’s sentiment during specific events such as stock market fluctuations, presidential elections, Oscars, tsunamis, or toward entities such as movies, companies, or aspects of a product (Bollen et al., 2011; Thelwall et al., 2011; Jiang et al., 2011; Hu and Liu, 2004; Jo and Oh, 2011).

To our knowledge, we are the first to explore recognition of affective polarity in similes as a whole, where the polarity relates to an act or state of the tenor. Unlike previous work, we do not rely on the presence of explicit properties. We also present a data set annotated with affective polarity in similes, and experiment with both manually annotated and automatically acquired training data.

### 3 Simile Data Set Creation

One of the major challenges of supervised classification is acquiring sufficient labeled data for training, since manual annotation is time consuming. However, similes sometimes contain words with explicit polarity (e.g., “bed feels like *heaven*” or “he *cries* like a baby”). Many of these cases can be identified with existing sentiment resources and then used to provide a classifier with training instances. But because sentiment resources have limitations (e.g., sentiment classifiers are not perfect, sentiment lexicons do not possess knowledge of context), these instances will have some noise. Therefore, we experiment with both manually labeled data sets that are smaller in size but high quality, and automatically labeled data sets that are comparatively larger but noisier.

Twitter is a popular microblogging platform and is widely used for sentiment analysis. Thus it is an excellent source for collecting similes that people use in everyday conversation. For this research, we extracted similes from 140 million tweets we harvested using the Twitter streaming API from March 2013 to April 2014. We started by selecting tweets containing three common **comparator** keywords: “*like*”, “*as*”, and “*than*”. We removed tweets with exact duplicate content, and tweets containing a retweet token. An additional challenge of tweets is that many are “near duplicates” (e.g., shared tweets with an added symbol or comment). So we performed an additional de-duplication step using Jaccard similarity of trigrams to measure overlap in the text content between pairs of tweets. When Jaccard similarity between two tweets was  $> 0.5$ , we kept only the

longer tweet, and repeated the process.

We used the UIUC Chunker (Punyakanok and Roth, 2001) to identify phrase sequences with the syntax of similes (e.g.,  $NP_1 + VP + PP\text{-like} + NP_2$ , or  $NP_1 + VP + ADJP + PP\text{-like} + NP_2$ , where  $NP_1$  is the tenor,  $NP_2$  is the vehicle, VP is the event, and ADJP is any explicitly mentioned property). We generalized over the extracted similes by removing the comparator, and the optional explicit property component. Our simile representation is thus a triple of the tenor, event and vehicle. We also lemmatized all words using Stanford CoreNLP (Manning et al., 2014). For a tenor phrase, we kept only the head noun, which is usually sufficient to understand the affective polarity target. We kept the entire noun phrase for the vehicle, since vehicles like “ice box” and “gift box” may represent two different concepts with different polarities in similes. We replaced personal pronouns (e.g., he, she) with a general PERSON token and other pronouns (e.g., it, this, that) with a general IT token. Table 2 presents examples of positive and negative similes in the annotated data set.

Positive	Negative
PERSON, smile, sun	PERSON, look, zombie
PERSON, feel, kid	PERSON, treat, stranger
PERSON, be, older brother	PERSON, feel, poo
IT, sound, heaven	PERSON, look, clown
PERSON, look, superman	word, cut, knife
IT, be, old time	PERSON, act, child
IT, feel, home	PERSON, look, voldemort
IT, fit, glove	PERSON, look, wet dog
IT, would be, dream	PERSON, treat, baby
IT, smell, spring	PERSON, look, drug addict

Table 2: Sample Similes from Annotated Data.

Sometimes, vehicle phrases contain adjective modifiers indicating explicit sentiment (e.g., “she looks like a *beautiful* model”). Since a simile is trivial to classify with such a modifier, we removed the instances that already had positive or negative adjective modifiers. To identify these cases, we used the AFINN sentiment lexicon (Nielsen, 2011). Similes that contain profanity (e.g., “You look like *crap*”) are nearly always negative, and trivial to classify, so we filtered out these cases using a freely available profanity list<sup>1</sup>. We also removed any simile where the vehicle is a pronoun (e.g., “it looks like that”), and discard similes appearing fewer than 5 times. Our final data set contains 7,594 similes.

<sup>1</sup><http://www.bannedwordlist.com/lists/swearWords.txt>

### 3.1 Manually Annotated Simile Data Set

To obtain manual annotation, we randomly selected 1500 similes occurring at least 10 times, from the 7,594 similes. Our expectation was that more frequent similes will be easier for the annotators to understand. We used Amazon’s Mechanical Turk to obtain gold standard annotations for affective polarity. We asked the annotators to determine if a simile expresses affective polarity toward the subject (i.e., the **tenor** component), and to assign one of four labels: *positive*, *negative*, *neutral*, or *invalid*. The first two labels are for similes that clearly express positive polarity (e.g., “*Jane swims like a dolphin*”) or negative polarity (e.g., “*Fred’s hair looks like a bird’s nest*”). The *neutral* label is for similes that do not have positive or negative polarity (e.g., “*the cloud looks like a turtle*” isn’t a positive or negative comment about the cloud) or similes that are ambiguous without the benefit of context (e.g., “*he is like my dog*” could be good or bad depending on the context).

The data also contained many misidentified similes, typically due to parsing errors. For example, sometimes there is an entire clause in place of the vehicle (e.g., “I feel like im gonna puke”). Other times, the informal text of Twitter makes the tweet hard to parse (e.g., “he is like whatttt”) or a verb occurs after “like” (e.g., “he is like hyperventilating”). The *invalid* label covers these types of erroneously extracted similes.

The annotation task was first conducted on a small sample of 50 similes, to select workers that had high annotation agreement with each other and gold standard labels we prepared. The best three workers then all annotated the official set of 1500 similes. The average Cohen’s Kappa ( $\kappa$ ) (Carletta, 1996) between each pair of annotators was 0.69. We then assigned the final label through majority vote. However, none of the annotators agreed on the same label for 78 of the 1500 similes, and 303 instances were labeled as *invalid* similes by the annotators. So we removed these 381 instances from the annotated data set. Finally, we randomly divided the remaining similes into an evaluation (Eval) set of 741 similes, and a development (Dev) set of 378 similes. Table 3 shows the label distribution of these sets.

### 3.2 Automatically Labeled Similes

For any new domain (e.g., Amazon product reviews), manual annotations for supervised training

Label	# of Similes (Dev Data)	# of Similes (Eval Data)
Positive	164	312
Negative	181	343
Neutral	33	86
Total	378	741

Table 3: Manually Annotated Data.

may not be readily available, and being able to automatically obtain training instances can be valuable. We therefore create and experiment with six types of automatically labeled training data.

**Using AFINN Sentiment Lexicon Words:** Our first training data set is created using the AFINN sentiment lexicon (Nielsen, 2011) containing 2,477 manually labeled words with integer values ranging from -5 (negativity) to 5 (positivity). For each simile, we sum the sentiment scores for all lexicon words in the simile components, assigning positive/negative polarity depending on whether the sum is positive/negative. This method yields 460 positive and 423 negative similes.

**Using MPQA Sentiment Lexicon Words:** Our second training data set is created using the 2,718 positive words and 4,910 negative words from the MPQA lexicon (Wilson et al., 2005). We applied the CMU part-of-speech tagger for tweets (Owoputi et al., 2013) to match the MPQA parts-of-speech for each word. We assign positive/negative polarity to similes with more positive/negative lexicon words. This method yields 629 positive and 522 negative similes.

**Using Sentiment Classifiers:** We create our third training data set using a state-of-the-art sentiment classifier designed for tweets. For this, we re-implemented the NRC Canada sentiment classifier (Zhu et al., 2014) using the same set of features described by the authors. We use a Java implementation<sup>2</sup> of SVM from LIBLINEAR (Fan et al., 2008), with the original parameter values used by the NRC Canada system. We trained the sentiment classifier with all of the tweet training data from SemEval 2013 subtask B (Nakov et al., 2013). We label a simile as positive or negative if the sentiment classifier labels it as positive or negative, respectively. This method yields 1185 positive and 402 negative similes.

**Using Sentiment in Surrounding Words:** The previous approaches for labeling training instances will primarily identify similes that contain one or more strongly affective words. This

<sup>2</sup><http://liblinear.bwaldvogel.de/>

can potentially bias the training data and limit the classifier’s ability to learn to recognize affective similes that do not contain words with a positive or negative connotation. Therefore, we explore an additional approach where instead of judging the sentiment of the words in the simile, we analyze the words in the tweet surrounding the simile. We hypothesize that there are often redundant sentiment indicators in the tweet. For example, “I *hate* it when my room is as cold as Antarctica”. For each simile, we identify all tweets that contain the simile and collect all of the words surrounding the simile in these tweets as a collective “context” for the tweet. We then count the number of distinct positive and negative sentiment words and compute the probability of positive or negative polarity given all the sentiment words surrounding a simile, and retain positive or negative similes with probability higher than a threshold (here, 0.7 to ensure high quality). As our sentiment lexicon, we combined the MPQA and the AFINN lexicon.

One issue is that when people feel amused (e.g., “he looks like a complete zombie, haha”) or sarcastic (e.g., “my room feels like an igloo. great! LOL.”), seemingly positive words in the context can be misleading because the sentiment is actually negative. As a simple measure to mitigate this issue, we manually removed a small set of laughter indicators from the lexicons (e.g., lol, haha).

This method yielded 492 positive and 181 negative similes.

**Combination of Training Instances:** As our last two training sets, we combined sets of instances labeled using the different methods above. As the fifth set, we combined training instances collected using the MPQA and AFINN lexicons and the NRC Canada sentiment classifier, which yielded a total of 2274 positive similes and 1347 negative similes. As our sixth set, we added the instances recognized from the surrounding words of a simile, producing the largest data set of 2766 positive and 1528 negative similes.

We also select neutral instances that are not identified as positive or negative by the above approaches and that also do not contain a sentiment lexicon (AFINN + MPQA) word in their collective context. For each approach, we then randomly select our final training instances for positive, negative and neutral classes maintaining the distribution of the development data. The final training data sizes are reported in Table 5.

## 4 Classifying Simile Polarity

Our goal is to create a classifier that can determine whether a simile expresses positive or negative affective polarity toward its subject. We present a classifier designed to label similes as Positive, Negative, or Neutral polarity. In this section, we describe the feature set and the classification framework of the supervised classifiers.

### 4.1 Feature Set

We extract three types of features from a simile, representing the lexical, semantic, and sentiment properties of the simile components.

#### 4.1.1 Lexical Features

**Unigrams:** A binary feature indicates the presence of a unigram in a simile. This feature is not component specific, so the unigram can be from any simile component (tenor, event or vehicle).

**Simile Components:** We define a binary feature for each tenor, event and vehicle phrase in the data set. This feature is component specific, (e.g., “dog” as a tenor is a different feature from “dog” as a vehicle).

**Paired Components:** We use a binary feature for each pair of simile components. Our intuition is that a pair of components may indicate affective polarity when used together. For example, “event:feel, vehicle:ice box” is negative for many different tenors (e.g., house, room, hotel). Similarly, “tenor:person, vehicle:snail” is negative for many different events (e.g., move, run, drive).

**Explicit Properties Associated with Vehicle:** Sometimes a simile explicitly mentions a property that is common to the tenor and the vehicle (e.g., “my pillow is *soft* like a cloud”). Although the properties are not part of our triples because they are optional components, we can still use them as valuable features, whenever present in the original corpus. For each simile vehicle, we therefore extract all explicit properties mentioned with that vehicle in our corpus, and create a binary feature for each (e.g., “Jane swims like a dolphin” and “Jim runs like a cheetah” can both share the feature *fast*, if *fast* appears with both “dolphin” and “cheetah” in the corpus as an explicit property).

**Vehicle Pre-modifiers:** We use a binary feature for each noun or adjective pre-modifier that appears with the vehicle (the vehicle head noun itself is excluded). Our intuition is that the same pre-modifiers appearing with different vehicles indi-

cate the same affective polarity (e.g., “smells like *wet* dog” and “smells like *wet* clothes”).

#### 4.1.2 Semantic Features

**Hypernym Class:** We obtain up to two levels of hypernym classes for each simile component head, using WordNet (Miller, 1995). For words with multiple senses, we only use the first synset of a word from WordNet, for simplicity. Once the hypernym classes are obtained for a word, we no longer keep the level information, and use a binary feature to represent each hypernym class. Our intuition is that groups of similar words can be used in different similes with the same affective polarity (e.g., room, bedroom).

**Perception Verb:** We create a binary feature to indicate if the event component is a perception verb. Perception verbs are fairly common in similes (e.g., “*looks* like a model”, “*smells* like garbage”). We use a set of the 5 most common perception verbs in similes (look, feel, sound, smell, taste).

#### 4.1.3 Sentiment Features

We add sentiment features that can be recognized in the simile using existing sentiment resources. For this purpose, we combined the MPQA (Wilson et al., 2005), and the AFINN lexicon (Nielsen, 2011) to use as our sentiment lexicon.

**Component Sentiment:** We use 3 binary features (one for each component) to indicate the presence of a positive sentiment word, and 3 binary features to indicate the presence of a negative sentiment word in each simile component.

**Explicit Property Sentiment:** We use 2 numeric features that count the number of positive and negative properties that appear with the vehicle in our corpus. We look for the property words in the combined AFINN and MPQA sentiment lexicons.

**Sentiment Classifier Label:** We use 2 binary features (one for positive and one for negative) to represent the label that the NRC-Canada Sentiment Classifier assigns to a simile.

**Simile Connotation Polarity:** We use 2 binary features (one for positive and one for negative) to indicate the overall connotation of a simile. We count whether the number of positive (or negative) connotation words is greater in a simile using a Connotation Lexicon (Feng et al., 2013), which contains 30,881 words with positive connotation and 33,724 words with negative connotation.

## 4.2 Classification Framework

As our supervised classification algorithm, we use a linear SVM classifier from LIBLINEAR (Fan et al., 2008), with its default parameter settings. Our goal is to assign one of three labels to a simile: *Positive*, *Negative*, or *Neutral*. We train two binary classifiers, one for positive and one for negative polarity. For positive polarity, we use similes labeled *positive* as positive training instances, and similes labeled *negative* or *neutral* as the negative training instances. For the negative polarity classifier, we use similes labeled *negative* as the positive training instances, and similes labeled *positive* or *neutral* as the negative instances.

To classify a simile, we apply both classifiers. If the simile is labeled as *positive* or *negative*, then it is assigned that label. If the simile is labeled as both *positive* and *negative*, or not labeled as either, then it is assigned a *neutral* label. We did not create a classifier to solely identify neutral similes because neutral similes are much less common than positive/negative similes, making up only 8.7% of the extracted similes in our development set (Table 3). Consequently, obtaining a large set of neutral similes via manual annotation would have required substantially more manual annotation effort. Secondly, we did not have a good way to reliably identify neutral similes automatically.

## 5 Evaluation

### 5.1 Classification Performance with Manually Annotated Data

Table 4 presents the results for supervised classification with our manually annotated data set using 10-fold cross-validation. As baselines, we used existing sentiment resources as described in Section 3.2, but now applied to evaluation data. We also used the connotation lexicon (Feng et al., 2013) the same way as the MPQA sentiment lexicon (Wilson et al., 2005) to compare as an additional baseline. The top section of Table 4 shows how effective these four existing sentiment resources are at assigning polarity to similes. Although precision was sometimes very high, recall was low across the board.

The lower section of Table 4 shows results for our classifiers. We first trained a classifier using only the sentiment features in order to shed light on the effectiveness of traditional sentiment indicators. Row (a) in Table 4 shows that this classifier produces reasonable precision (65-72%) but recall

	Positive			Negative			Neutral		
	P	R	F	P	R	F	P	R	F
<i>Sentiment Resource Baselines</i>									
AFINN Lexicon	<b>88</b>	17	28	<b>95</b>	18	31	13	<b>95</b>	23
MPQA Lexicon	83	21	34	90	15	26	13	<b>95</b>	24
Connotation Lexicon	61	38	47	63	40	49	17	63	26
NRC Canada Sentiment Classifier	72	34	47	94	16	27	13	83	23
<i>Affective Polarity Simile Classifiers</i>									
(a) Sentiment Features	65	54	59	72	48	58	19	37	25
(b) Unigrams	73	52	61	74	70	72	21	47	29
(c) Unigrams + Other Lexical	73	56	63	75	76	75	<b>26</b>	<b>45</b>	<b>33</b>
(d) Unigrams + Other Lexical + Semantic	68	59	63	76	72	74	24	40	30
(e) Unigrams + Other Lexical + Semantic + Sentiment	75	<b>60</b>	<b>67</b>	77	<b>79</b>	<b>78</b>	25	40	31

Table 4: Results with Manually Annotated Training Data (P = Precision, R = Recall, F = F1-score).

Classifier	# of Training Instances			Positive			Negative			Neutral		
	Pos	Neg	Neu	P	R	F	P	R	F	P	R	F
(a) SVM with labeled data using AFINN	384	423	78	<b>78</b>	32	45	85	31	45	<b>14</b>	80	<b>24</b>
(b) SVM with labeled data using MPQA	475	522	94	65	44	53	81	27	41	12	59	20
(c) SVM with labeled data using NRC Canada	365	402	74	72	34	47	<b>94</b>	16	27	13	<b>83</b>	23
(d) SVM with labeled data from (a), (b), + (c)	1085	1193	216	69	50	58	<b>88</b>	30	45	13	62	22
(e) SVM with labeled data using sentiment in surrounding words	164	181	34	60	57	59	62	<b>57</b>	<b>60</b>	13	20	16
(f) SVM with labeled data from (a), (b), (c), + (e)	1221	1342	242	64	<b>61</b>	<b>62</b>	75	48	59	11	30	16

Table 5: Results with Automatically Labeled Training Data (P = Precision, R = Recall, F = F1-score).

levels only around 50% for both positive and negative polarity. The Neutral class has extremely low precision, which indicates that many unrecognized positive and negative similes are being classified as Neutral.

Row (b) shows the results for a baseline classifier trained only with unigram features. Unigrams perform substantially better than the sentiment features for negative polarity, but only slightly better for positive polarity. Row (c) shows that the additional lexical features described in Section 4.1.1 further improve performance.

Row (d) shows that adding the semantic features did not improve performance. One reason could be that some WordNet hypernym classes are very specific and may not generalize well. Also, similes can have different polarities with vehicle words from the same general semantic class (e.g., “he runs like a *cheetah*” vs “he runs like a *turtle*”).

Finally, Row (e) shows that adding the sentiment features along with all the other features yields a precision gain for positive polarity and a recall gain for negative polarity. Overall, the full feature set improves the F score from 61% to 67% for positive polarity, and from 72% to 78% for negative polarity, over the unigram baseline.

## 5.2 Classification Performance with Automatically Acquired Training Data

Table 5 shows the performance of the classifiers (using our full feature set) when they are trained with automatically acquired training instances. The upper section of Table 5 shows results using training instances labeled by three different sentiment resources. Row (d) shows that combining the training instances labeled by all three resources produces the best results.

Row (e) of Table 5 shows the performance of the classifiers when they are trained with instances selected by analyzing sentiment in the surrounding words of the similes. We observe a substantial recall gain, which validates our hypothesis that similes obtained by recognizing sentiment in their surrounding words provide the classifier with a more diverse set of training examples. Finally, Row (f) shows that using both types of training instances further improves performance for positive polarity, and increases precision for negative polarity but with some loss of recall.

Comparing these results with Table 4, we see that there is still a gap between the performance of classifiers trained with manually annotated data versus automatically acquired data. However, the classifiers trained with automatically acquired data produce substantially higher F scores than all of

the baseline systems in Table 4. Using automatically acquired training data is a practical approach for creating simile classifiers for specific domains, such as Amazon product reviews (e.g., “headphone sounds like garbage”, or “each song is like a snow-flake”) which were studied in previous work on figurative comparisons in similes (Niculae and Danescu-Niculescu-Mizil, 2014).

### 5.3 Impact of Training Data Size

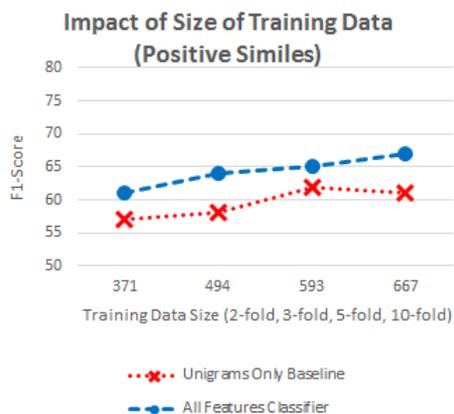


Figure 1: Learning Curve for Positive Similes.

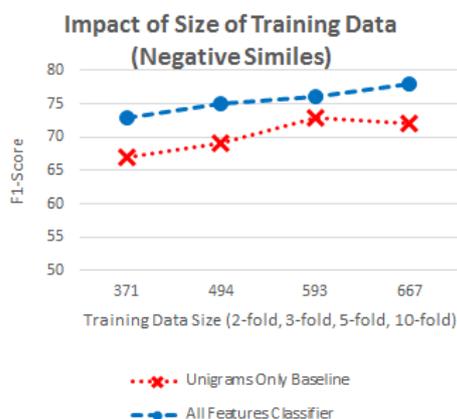


Figure 2: Learning Curve for Negative Similes.

We also generated learning curves to determine how much the size of the training set matters. Figures 1 and 2 show the performance of classifiers trained using varying amounts of manually annotated data. We show results for the classifiers trained only with unigram features and classifiers trained with our full feature set, for positive similes in Figure 1 and negative similes in Figure 2. The results were produced from 2-fold, 3-fold, 5-fold and 10-fold cross-validation experiments, with the size of the corresponding training sets shown on the X-axis. These figures show that

the classifiers with unigram features hit a plateau at about 600 training instances. However the classifiers with the full feature set continually benefited from more training data. Table 6 presents a sample of similes where the vehicle appears only once in our data set. The unigram-based classifier could not classify these instances, but the classifier with the full feature set could.

Positive	Negative
PERSON, feel, superhero	PERSON, feel, old woman
PERSON, be, friend	PERSON, be, hurricane
beast, look, beauty	IT, feel, eternity
PERSON, feel, hero	PERSON, feel, peasant
PERSON, feel, champion	PERSON, eat, savage
PERSON, seem, sweetheart	PERSON, be, witch
IT, be, sleepover	PERSON, feel, prisoner
IT, be, reunion	IT, be, north pole
PERSON, feel, president	IT, feel, winter
ronaldo, be, messi	PERSON, be, wolf

Table 6: Similes with unique vehicles that were correctly classified using the full feature set.

## 6 Analysis and Discussion

We also conducted a qualitative analysis of our new corpus of similes and the behavior of the classifiers. We hypothesized that there are at least two reasons why similes might be difficult to classify. First, the interpretation of a simile can be highly context-dependent and subjective, depending on the speaker or the perceiver. To illustrate, Table 7 presents examples of similes that can have different polarity depending on the speaker or perceiver’s personal experience or location, and other subjective aspects of the context. For example, *it looks like snow* may be a good thing to someone who lives in Utah where people look forward to skiing, but a bad thing to someone living in Boston during the winter of 2015. Similarly *she smells like a baby* is typically positive to new mothers, but was viewed as negative by the Mechanical Turk annotators.

Polarity	Simile	Context
positive	PERSON, smell, baby	young mother
negative	PERSON, smell, baby	MTurkers
negative	IT, look, snow	lives in Boston
positive	IT, look, snow	lives in Utah
negative	IT, look, rain	lives in England
positive	IT, look, rain	lives in California

Table 7: Similes with Context-dependent Polarity.

Second, we hypothesized that the polarity of a simile might interact with the distinction made in previous work between figurative and literal

uses of similes (Bredin, 1998; Addison, 1993), for example Niculae and Danescu-Niculescu-Mizil (2014) showed that sentiment and figurative comparisons are strongly correlated. Thus our expectation was that most literal comparisons would be neutral while most figurative comparisons would carry polarity. To explore this issue, we conducted an informal analysis of the 378 similes in our development data set to examine the literal vs. figurative distinction. For this analysis, we looked at the simile component triples as well as the context of ten tweets in which the simile appeared.

Our analysis suggests that the picture is more complex than we initially hypothesized. We found that, 1) the distinction between positive and negative similes in our data is orthogonal to the figurative vs. literal distinction, 2) some similes are used both figuratively and literally, and cannot be differentiated without context, 3) even in cases when all sample uses were literal, it is easy to invent contexts where the simile might be used figuratively, and vice versa, and 4) for a particular instance (simile + context), it is usually possible to tell whether a figurative or literal use is intended by examining the simile context, but some cases remain ambiguous. Table 8 shows examples of some similes that we identified as being figurative, literal, or both depending on context.

Use	Polarity	Simile
fig	positive	house, smell, heaven
fig	positive	PERSON, look, queen
fig	negative	PERSON, look, tomato
lit	negative	hair, smell, smoke
lit	neutral	PERSON, look, each other
both	neutral	house, smell, pizza
both	negative	IT, smell, skunk
both	negative	PERSON, look, frankenstein

Table 8: Similes with figurative (fig) or literal (lit) interpretation, or ambiguous depending on the context.

These observations reinforce the difficulty with making the figurative/literal distinction noted by Niculae and Danescu-Niculescu-Mizil (2014), whose annotation task required Turkers to label comparisons on a scale of 1 to 4 ranging from very literal to very figurative. Even with Master Turkers, a qualification task, filtering annotators by gold standard items, and collapsing scalar 1,2 values to literal and 3,4 values to figurative, the inter-annotator agreement with Fleiss’  $\kappa$  was 0.54. They note that out of 2400 automatically extracted comparison candidates, only 12% end up being se-

lected confidently as figurative comparisons.

Selected cases that the classifiers fail on are further illustrated in Table 9. Examples S1 to S9 could be related to the difficulties noted above with subjectivity of interpretation. Many people for example like the smell of coffee and pizza, but perhaps not when a person smells that way. Similarly, *a baby* is often positive as a vehicle, but smelling and sounding like a baby may not be positive depending on the circumstances, while the positive or negative interpretation of *sounding like a pirate* and *looking like a pirate* might also be context dependent.

ID	Simile	Gold	Man	Auto
S1	PERSON, smell, coffee	neg	pos	pos
S2	PERSON, smell, pizza	pos	neg	neut
S3	IT, smell, pizza	neut	pos	neut
S4	PERSON, sleep, baby	pos	pos	neut
S5	PERSON, smell, baby	neg	neg	neut
S6	PERSON, feel, baby	pos	neg	pos
S7	PERSON, sound, baby	neg	pos	neut
S8	PERSON, sound, pirate	pos	neg	neg
S9	PERSON, look, pirate	neg	neg	neut

Table 9: Error Analysis of Classifier Output (Man = Classifier trained with manually annotated instances, Auto = Classifier trained with automatically annotated instances).

## 7 Conclusions

Similes are one example of a tractable case of sentiment-bearing expressions that are not recognized well by current sentiment analysis tools or lexicons. Making progress on sentiment analysis may require tackling many different types of linguistic phenomena such as this one. To this end, we have presented a simile data set labeled with affective polarity and have presented a supervised classification framework for recognizing affective polarity in similes. We have also presented our experiments with both manually labeled and automatically acquired training instances. We have shown that with manually labeled data, our feature set can substantially improve performance over a unigram only baseline. We have also shown that good performance can be achieved with automatically acquired training instances, when manually labeled data may not be available.

## Acknowledgments

This material is based upon work supported in part by the National Science Foundation under grants IIS-1450527 and IIS-1302668.

## References

- Catherine Addison. 1993. From literal to figurative: An introduction to the study of simile. *College English*, pages 402–419.
- Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.
- Hugh Bredin. 1998. Comparisons and similes. *Lingua*, 105(1):67–78.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.
- Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria, August. Association for Computational Linguistics.
- David Fishelov. 2007. Shall i compare thee? simile understanding and semantic categories. *Journal of literary semantics*, 36(1):71–87.
- Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado, June. Association for Computational Linguistics.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.
- Patrick Hanks. 2005. Similes and sets: The english preposition like. *Languages and Linguistics: Festschrift for Fr. Cermak*. Charles University, Prague.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.
- Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.
- Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.
- Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541.
- Bin Li, Haibo Kuang, Yingjie Zhang, Jiajun Chen, and Xuri Tang. 2012. Using similes to extract basic sentiments across languages. In *Web Information Systems and Mining*, pages 536–542. Springer.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.
- Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.
- Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018. Association for Computational Linguistics.
- Finn Årup Nielsen. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs",. In *ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages*.

- Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.
- Anthony M Paul. 1970. Figurative language. *Philosophy & Rhetoric*, pages 225–248.
- V. Punyakanok and D. Roth. 2001. The use of classifiers in sequential inference. In *NIPS*, pages 995–1001. MIT Press.
- Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.
- Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and A. George Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375. Association for Computational Linguistics.
- Tim Rumbell, John Barnden, Mark Lee, and Alan Wallington. 2008. Affect in metaphor: Developments with wordnet. In *Proceedings of the AISB Convention on Communication, Interaction and Social Intelligence*, volume 1, page 21.
- Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.
- Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of CogSci*.
- Tony Veale. 2012. A context-sensitive, multi-faceted model of lexico-conceptual affect. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 75–79. Association for Computational Linguistics.
- Alan Wallington, Rodrigo Agerri, John Barnden, Mark Lee, and Tim Rumbell. 2011. Affect transfer by metaphor for an intelligent conversational agent. In *Affective Computing and Sentiment Analysis*, pages 53–66. Springer.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. 2014. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.